DOCUMENT RESUME

ED 029 699 PS 001 819

By-Larkins, A. Guy; Shaver, James P.

Matched-Pair Scoring Technique Used on a First-Grade Yes-No Type Economics Achievement Test.

Pub Date 21 Apr 67

Note-17p.; Paper presented to the Education Section of the Utah Academy of Sciences, Arts, and Letters, Logan, Utah, April 21, 1967

EDRS Price MF-$0.25 HC-$0.95

Descriptors-*Achievement Tests, Economics, *Grade 1, Primary Grades, Scoring, *Test Construction, Testing, *Testing Problems, Test Reliability

Identifiers-Families At Work, *Matched Pair Scoring

There exist special problems in testing first-grade children. Orally administered yes-no tests reduce the problems found in the other types, but they have their own drawbacks. A solution to some of these drawbacks is the use of the matched-pair scoring technique. For each "yes" item on the test there is included a "reversed" or "no" item on the same concept being tested, and vice versa. The pupil must respond correctly to both in order to be given credit for either one. However, the drawback then becomes the necessity of doubling the size of the test. A 30-item test, based on "families at work" economics, was administered to six first grade classes; three had bee n studying the material, and three had not. The test results showed that (1) the matched-pair scoring technique increased the reliability of the yes-no test and also increased and general discriminatory power of the test and (2) the students who were studying the economics material scored higher than students who were not, indicating that the program was resulting in demonstrable learning. Another article by the same authors (see PS 001 822) also deals with this subject. (WD)

MATCHED-PAIR SCORING TECHNIQUE USED ON A FIRST-GRADE YES-NO TYPE ECONOMICS ACHIEVEMENT TEST*

A. Guy Larkins
and
James P. Shaver
Utah State University

PROBLEM

Testing first-grade children poses special problems, many of which are related to the limited reading ability of six-year-olds. A common approach to this problem is to use picture-type multiple-choice tests. However, the production of such tests is a formidable task for classroom teachers and/or small-scale research projects. For instance, a fifty-item five-option multiple-choice test requires two-hundred and fifty pictures. The YES-NO response test, presented orally by the tester, is a tempting alternative, but it also presents some difficulties, not the least of which are low reliability and pronounced response set. (Barnes, 1962; Cronbach, 1942, 1946, and 1950)

YES-NO type tests present the subject with only two options. Since a subject may respond randomly to such a test and still be correct half of the time, YES-NO tests tend to have low reliability unless most subjects are knowledgable concerning the content of the test and respond correctly to most of the items. Difficulties of test interpretation in YES-NO tests are further compounded by acquiescence-set: i.e., the tendency of students to respond YES when in doubt. A minority of students exhibit the opposite of acquiescence and respond NO when in doubt. (This latter response-set will be called "dissent set" in the remainder of this paper.)

---

* Paper presented to the Education Section of the Utah Academy of Sciences, Arts, and Letters. Logan, Utah, April 21, 1967.

In three investigations of students' responses to first-grade economics achievement tests, the authors found that frequency of correct response to items for which the correct response is YES (hereafter called YES items) was approximately 70% while the frequency of correct response to items for which the correct response is NO (hereafter called NO items) was approximately 40%. (The theoretical expected frequency of correct response for both types of items if students were responding randomly is 50%.)

Since most students exhibit acquiescence rather than dissent, the frequency of correct responses to YES items may be spuriously high. An example might illustrate how this can affect the interpretation of test results if the tester is not aware of the problem. In an investigation by Shaver and Larkins (1966), control and experimental groups each containing approximately 100 first-grade children were asked to respond to this statement, "A specialist is a man who learns to do one job very well." Ninety children in each group correctly responded YES. Since this frequency of correct response is clearly above the theoretical expectation of 50, we might be tempted to conclude that children in both the control and experimental groups knew the concept being tested. However, when the same children were asked to respond to, "A specialist can do more things for himself than a person who has not specialized," approximately thirty-five children in each group correctly responded NO. This is significantly below the theoretical expected frequency of fifty and would seem to indicate that most of the children did not know the meaning of the word "specialist," thus contradicting the result on the previous question. It is possible that the majority of the ninety correct responses to the YES item were due to acquiescence rather than knowledge.

In brief, interpretation of individual YES-NO test items is difficult since there is no way of determining what portion of the responses is due to acquiescence-dissent, and what portion is due to knowledge. Of course, interpretation of scores of individual students is always difficult when tests suffer from low reliability.

A technique, which hereafter will be referred to as "matched-pair scoring," was devised to cope with both of the above problems. Matched-pair scoring involves writing reversed items for each concept or bit of information tested. "Reversed items" means that for every YES item there is a NO item intended to test the same content. For example,

CHILDREN WHO JUMP ROPE ARE PRODUCERS. (NO)

CHILDREN WHO WASH DISHES ARE PRODUCERS. (YES)

In matched-pair scoring the students are required to correctly respond to both forms of an item before credit is given for either. Therefore, if students are responding from acquiescence they will incorrectly respond to the NO items. If students are responding from dissent, they will incorrectly respond to the YES items. A correct response to both items indicates either knowledge or an occasional lucky guess (probability of one in four) on both items.

The literature on acquiescence contains several references to attempts to write reversed items for personality inventories. (Mogar, 1960; Chapman and Campbell, 1957; Bass, 1955; Leavitt, Hax, and Roche, 1955; Rokeach, 1963; Christie, Havel, and Seidenberg, 1958; Peabody, 1961; Rorer, 1963; and Ong, 1963.) Most of these discussions are of little value for this project due to differences between writing items designed to test knowledge and writing items designed to measure personality traits. A review

of the literature on acquiescence did not uncover any attempts to write reversals for achievement tests with the intention of improving the reliability of the test by correcting for response-set.

Cronbach (1942) has suggested that the reliability of YES-NO tests can be increased by writing tests containing only NO items. Since most people tend to acquiesce rather than dissent, a NO response would generally be made from knowledge. However, NO-item-only tests favor the dissenter. A person who tends to respond NO will have a spuriously high score on such a test. The matched-pair scoring technique should cancel both the effects of acquiescence and dissent.

One drawback of the matched-pair technique is that it reduced the size of the test by half. A sixty-item test becomes a thirty-item test since pairs of items are scored as one. The positive effect on reliability of increasing the options on each item from two to four must outweigh the negative effect of halving the length of the test in order for this technique to be useful.

The problem, then, is that the various test formats available for primary-grade teacher-made tests either require reading ability on the part of the child (written multiple-choice tests), require too much time to construct (picture-type multiple-choice tests), or suffer from low reliability and/or are subject to acquiescence-dissent set (YES-NO tests). Orally administered YES-NO tests require a minimum of reading ability, and compared to picture-type tests, can be constructed in a reasonable amount of time. Their usefulness is limited, however, unless a way can be found to increase their reliability and correct for response-set.

OBJECTIVES

The objectives of this study were to:

1. Devise a YES-NO type first-grade economics achievement test using matched pairs of reversed items, and

2. Determine whether this technique would increase the reliability of the test.

Since the content of the achievement test was based on OUR WORKING WORLD: FAMILIES AT WORK (Senesh, 1964), a secondary objective was to determine if a group of first-grade children could learn certain concepts presented in Lessons Four and Five of that course of study.

PROCEDURE

Two college students* in an elementary education course taught by one of the authors selected key terms and concepts from Lessons Four and Five of FAMILIES AT WORK, and wrote a thirty item YES-NO test with reversals--i.e., thirty items, making fifteen matched pairs. The test was deliberately kept short in length since it was used in the early part of the first grade when children have had little experience with test procedures.

In writing items, one special precaution was taken. No YES item included the words "no" or "not." Just prior to the construction of this test, a similar test had been administered to one-hundred and fifty-six first-grade children. One of the teachers suggested that children seemed confused by the word "no" in some items. The children appeared to be answering items containing "no" in just the opposite manner from that intended, so that they answered NO when they intended to answer YES, and vice versa.

---

* The authors are indebted to Dora Buckhouse and Clare Bowen for substantial assistance during the course of this project.

To test this suggestion, items were divided into four categories:

YES items containing "no" (hereafter called YESno)
YES items not containing "no" (hereafter called YES)

NO items containing "no" (hereafter called NOno)
NO items not containing "no" (hereafter called NO)

Knowing that acquiescence-dissent causes different patterns of correct response to YES and NO items (see page 2), we hypothesized that:

> If the word "no" is not causing children to respond opposite to their intention, the frequency of correct response will not significantly differ between the categories YESno and YES, or between NOno and NO, but the two YES categories will differ from the two NO categories.

In other words, if the word "no" causes children to respond No to YES items and Yes to NO items, then the YESno items will produce a pattern of correct response similar to that produced by NO items, and the NOno items will produce a pattern of correct response similar to that produced by YES items.

Table One

Mean Frequency of Correct Response
to Four Categories of YES and NO Items

	Number of Items	Mean Frequency	
YES items not containing "no"	7	106	To reject our hypothesis, the frequencies for these two groups would have to be similar.
YES items containing "no"	7	75	
NO items not containing "no"	10	76	
NO items containing "no"	6	73	

Analysis of variance was used to test the significance of the differences between the four categories in Table One. The only significant difference was between YES and the other three categories.* The categories NO, NOno, and YESno did not differ significantly from each other. Part of our hypothesis is thus supported. It appears that the word "no" in YES items confuses children, but does not do so in NO items. Therefore, in constructing the present test we avoided including "no" in the YES items.

In November, 1966, the thirty-item YES-NO test was administered to six classrooms of first-grade children (three classes who had recently been taught Lessons Four and Five of FAMILIES AT WORK and three classes who had not) In two schools in separate districts in northern Utah. Testing soon after the completion of the lessons was done deliberately in order to maximize the students' chances of responding correctly since the content seemed quite difficult for young children. Control and experimental groups were tested the same morning. The control group was tested first. Each of the three persons involved in devising the test administered it to one control and one experimental class.

Tests were first corrected in the ordinary manner and then again using the matched-pair technique. Spilt-half reliability coefficients were computed for scores based on both scoring techniques to test the hypothesis that reliability would increase when the matched-pair method was used. Means and standard deviations were computed, and the t-test was used to compare the achievement of control and experimental groups.

---

*Significant beyond the .01 level.

Results

Table Two

Split-half Reliability:
Comparison of Control and Experimental Groups Using
Ordinary and Matched-Pair Scoring

	r_{11} ordinary	r_{11} matched
Experimental	.35	.60
Control	.14	.46

Expectations in regard to reliability were supported. Reliability for both control and experimental groups increased using matched-pair scoring. Under both scoring methods, reliability was greater for the experimental group, which was to be expected since the control students had not studied the material upon which the test was based and probably responded either randomly or from acquiescence-dissent on most items. A reliability coefficient of .60 for the experimental group is not high, yet is probably as high as one might reasonably expect for a fifteen-item test. The r_{11} for a test twice as long would be .75, estimated with the Spearman-Brown Prophecy Formula.*

Related to the primary objective of this project--determining whether matched-pair scoring increases reliability--is the question, "Does matched-pair scoring increase the discrimination power of the test?" Notice in Table Three that the standard deviations and t-ratios are larger for matched-pair scoring. The difference between standard deviations was tested

* Subsequent testing of the same group with a similar instrument containing twice as many items produced a split-half reliability coefficient of .75.

using the standard error of the standard deviation for correlated groups (Garrett, 1958, pp. 233-234) and in each case found to be significant beyond the .01 level.* Both the increase in standard deviations and the increase in t-ratios indicate that this test can be expected to discriminate better when scored using matched-pairs. The reader may also notice in Table Three that the difference between means increased.

Table Three

Comparison of Control and Experimental Groups Using Ordinary and Matched-Pair Scoring

	Ordinary Scoring				Matched-Pair Scoring			
	$\bar{X}$*	Diff.****	SD**	t-ratio***	$\bar{X}$	Diff	SD	t-ratio
Experimental Group	18.49	1.50	3.18	3.33	9.76	2.69	4.97	3.83
Control Group	16.99		2.34		7.07		3.86	

* $\bar{X}$ = Means
** SD = Standard deviation
*** With degrees of freedom = 70, a t-ratio of 2.65 is needed to be significant at the .01 level.
**** Diff = difference between means.

Table Three also shows that the experimental group scored significantly higher than the control group, indicating that they learned at least some of the content of Lessons Four and Five of FAMILIES AT WORK.

In order to determine which items accounted for the mean difference between groups, an item analysis was performed using Chi-square contingency tables to compare frequency of correct response on individual items.

---

* There is some question about the appropriateness of this procedure since the difference between SD's was not based on two separate testings, but on two methods of scoring a test which was given only once.

Table Four

Items Which Discriminated Between Groups

Item #	type	ORDINARY SCORING Frequency Control	Experimental	χ^2	MATCHED-PAIR SCORING Frequency Control	Experimental	χ^2
16	NO	38	55	7.73**	31	44	3.80*
26	YES	60	63	NS***			
5	NO	36	49	3.77	19	35	6.72**
6	YES	52	62	2.54			
1	NO	24	35	2.90	16	32	7.19**
9	YES	60	70	3.65			
11	NO	27	39	3.32	12	25	5.59*
27	YES	54	56	NS			
29	NO	12	35	15.53**	8	30	16.22**
3	YES	60	66	NS			
23	NO	14	21	1.59	10	20	3.83*
20	YES	66	70	NS			

* 3.81 is significant at the .05 level of confidence.
** 6.64 is significant at the .01 level of confidence.
*** NS = not significant by inspection.

Using ordinary scoring, three of thirty items discriminated between groups at or near the .05 level or better. All three were NO items, which lends some support to Cronbach's contention that a NO-item-only test will discriminate between groups nearly as well as a YES-NO test which is twice as long. While only three of thirty items discriminated using ordinary scoring, six of fifteen pairs of items discriminated, which helps to explain the increased difference between means, increased standard deviations, larger t-ratios, and increased reliability when matched-pair scoring was used.

The reader may also be interested to note in Table Four that the frequency of correct response to NO items is consistently lower than for YES items, indicating that acquiescence-set is more powerful than dissent-set. Note also that the loss of frequency of correct response to YES items is generally greater for the control group than for the experimental group when matched-pair scoring is used. For example, on Item 27 the frequency of correct response for control and experimental groups is 54 and 56 respectively using ordinary scoring. When matched-pair scoring is used this drops to 12 and 25--a loss of 40 for the control group and 31 for the experimental group. It would seem that students in the control group responded more often from acquiescence, and that ordinary scoring obscured the real difference between groups.

The loss in frequency of correct response was not limited to YES items. The frequency of correct response to every NO item in Table Four decreased under matched-pair scoring. For instance, Item 5 decreased from 36 to 19, a loss of 17 correct responses. This indicates that some responses are made either randomly or from dissent rather than acquiescence, and tends to support our reluctance to use NO-item-only tests.

CONCLUSION

The purpose of this study was to determine whether matched-pair scoring would increase the reliability of YES-NO tests administered to young children. Findings were positive. Not only did reliability increase, but so did the general discriminatory power of the test. Our secondary purpose was to determine whether children could learn the terms and concepts found in Lessons Four and Five of FAMILIES AT WORK. Again, findings were positive. Mean differences between control and experimental groups were significant beyond the .01 level.

Several words of caution should be added to the generally optimistic tone of this paper. First, this project lacked rigorous design controls. Children were not assigned randomly to treatments (a difficult criterion to meet in many investigations in education), nor were they selected randomly. This study was performed as an exploratory investigation, and its limitations should be kept in mind.

Secondly, we have experienced some success with matched-pair scoring. The reliability obtained ($r_{11} = .75$) is acceptable for research involving fairly large samples, but may not be adequate to the demands of classroom assessment of individuals.

For those who are especially interested in economic education for young children, especially with the FAMILIES AT WORK materials, the complete test used in this project is presented in the following appendix. Items are listed in pairs in order of their ability to discriminate between experimental and control groups. Not all of the items which do not discriminate at the .05 level or better should be excluded from further testing. The frequency of correct response on most items is larger for the experimental group, though not significantly so, and thus accumulate to increase the discriminatory power of the test in general. Also, some items which do not discriminate should be included for content validity. They may indicate inability of children to learn certain concepts, rather than indicate poor item construction.

APPENDIX

PRIMARY ECONOMICS TEST: LESSONS FOUR AND FIVE

OF FAMILIES AT WORK

X^2 stands for chi-square.
X^2M stands for chi-square using matched-pair scoring technique.
FC stands for frequency of correct response in the control group.
FE stands for frequency of correct response in the experimental group.
FEM or FCM stands for frequency of correct response in control or experimental groups when using match-pair scoring.
INSP stands for not significant by inspection.
Levels of significance for chi-square are 3.81 at the .05 level and 6.64 at the .01 level.

Item #

29. WHEN MOTHER AND SISTER WATCH T.V. THEY ARE DIVIDING THE LABOR. (NO)
FC=12, FE=35, X^2=15.43

3. WHEN MOTHER WASHES THE DISHES AND SISTER DRIES THEM THEY ARE DIVIDING THE LABOR. (YES)
FC=60, FE=66, X^2=insp

FCM=8, FEM=30, X^2M=16.22

1. A FARMER WHO RAISES WEEDS IS A PRODUCER OF GOODS. (NO)
FC-24, FE=35, X^2=2.90

9. A FARMER WHO RAISES POTATOES IS A PRODUCER OF GOODS. (YES)
FC=60, FE=70, X^2=3.65

FCM=16, FEM=32, X^2M=7.19

5. CHILDREN WHO JUMP ROPE ARE PRODUCERS. (NO)
FC=36, FE=49, X^2=3.77

6. CHILDREN WHO WASH DISHES ARE PRODUCERS. (YES)
FC=52, FE=62, X^2=2.54

FCM=19, FEM=35, X^2M=6.72

11. IT IS FASTER AND CHEAPER FOR ONE MAN TO PRODUCE ALL OF HIS OWN GOODS. (NO)
FC=27, FE=39, X^2=3.32

27. IT IS FASTER AND CHEAPER TO DIVIDE THE LABOR. (YES)
FC=54, FE=56, X^2= insp

FCM=12, FEM=25, X^2M=5.59

23. WHEN TWO BABIES ARE PLAYING WITH DOLLS THEY ARE DIVIDING THE LABOR. (NO)
FC=14, FE=21, X^2=1.59

20. WHEN BROTHER SWEEPS THE FLOOR AND SISTER MAKES THE BED THEY HAVE DIVIDED THE LABOR. (YES)
FC=66, FE=70, X^2=insp

FCM=10, FEM=20, X^2M=3.83

16. WHEN PEOPLE SHOVEL SNOW ONTO THE SIDEWALK THEY ARE PRODUCING A SERVICE. (NO)
FC=38, FE=55, X^2=7.73

26. WHEN PEOPLE SHOVEL SNOW OFF THE SIDEWALK THEY ARE PRODUCING A SERVICE. (YES)
FC=60, FE=63, X^2=insp

FCM=31, FEM=44, X^2M=3.8

17. EVERYONE EXCEPT BABIES AND SICK PEOPLE ARE PRODUCERS. (YES)
FC=42, FE=47, X^2=insp

7. EVERYONE IS A PRODUCER. (NO)
FC=42, FE=45, X^2=insp

FCM=22, FEM=30, X^2M=1.42

25. MOTHER AND FATHER ARE THE ONLY CONSUMERS IN THE FAMILY. (NO)
FC=33, FE=45, X^2=3.17

30. EVERYONE IN THE FAMILY IS A CONSUMER. (YES)
FC=49, FE=50, X^2=insp

FCM=22, FEM=29, X^2M=1.18

4. PEOPLE WHO MAKE USEFUL THINGS ARE PRODUCERS OF SERVICES. (NO)
FC=15, FE=19, X^2=insp

15. PEOPLE WHO MAKE USEFUL THINGS ARE PRODUCERS OF GOODS. (YES)
FC=57, FE=66, X^2=2.32

FCM=12, FEM=15, X^2M=insp

2. IN THE FIRST-GRADE WHEN EACH CHILD GETS HIS OWN PAPER AND PENCIL THE CHILDREN HAVE DIVIDED THE LABOR. (NO)
FC=19, FE=22, X^2-insp

18. IN THE FIRST-GRADE WHEN ONE CHILD PASSES OUT PAPER AND ANOTHER CHILD PASSES OUT PENCILS THEY HAVE DIVIDED THE LABOR. (YES)
FC=66, FE=71, X^2=insp

FCM=12, FEM=17, X^2M=.91

10. WHEN THE DOG GUARDS THE HOUSE AND THE CAT CATCHES MICE THEY ARE DIVIDING THE LABOR. (YES)
FC=59, FE=55, X^2=insp

12. WHEN THE DOG AND CAT EAT THEIR DINNERS THEY ARE DIVIDING THE LABOR. (NO)
FC=15, FE=20, X^2=.77

FCM=9, FEM=13, X^2M=insp

21. WHEN FATHER HELPS JOHNNY BUILD A BOAT HE IS A PRODUCER OF GOODS. (YES)
FC=66, FE=63, X^2=insp

13. WHEN FATHER SPANKS JOANIE FATHER IS A PRODUCER OF GOODS. (NO)
FC=39, FE=45, X^2=insp

FCM=33, FEM=35, X^2M=insp

24. NATIONS WHO TRADE WITH EACH OTHER DIVIDE THE LABOR. (YES)
FC=55, FE=51, X^2=insp

22. NATIONS WHO TRADE WITH EACH OTHER DO NOT DIVIDE THE LABOR. (NO)
FC=41, FE=48, X^2=insp (but probably around 1.0)

FCM=27, FEM=26, X^2M=insp

14. A SPECIALIST DEPENDS ON OTHERS TO PRODUCE THE THINGS HE NEEDS. (YES)
FC=59, FE=56, X^2=insp

19. A SPECIALIST PRODUCES EVERYTHING HE NEEDS. (NO)
FC=25, FE=21, X^2=insp

FCM=18, FEM=15, X^2M=insp

8. WHEN MOTHER BAKES A CAKE SHE IS A PRODUCER OF GOODS. (YES)
FC=68, FE=72, X^2=insp

28. WHEN MOTHER CUTS JOHNNY'S HAIR SHE IS A PRODUCER OF GOODS. (NO)
FC=16, FE=15, X^2=insp

FCM=12, FEM=11, X^2M=insp

BIBLIOGRAPHY

Barnes, Eugene. MTA Conference 1962. An unpublished paper given at the Fourth Annual Military Testing Association Conference, October 22-25, 1962.

Bass, Bernard M. "Authoritarianism or Acquiescence?" Journal of Abnormal and Social Psychology, 1955, 51, 616-622.

Chapman, Loren J. and Donald T. Campbell. "Response Set in the F Scale." Journal of Abnormal and Social Psychology, 1957, 54, 129-134.

Christie, Richard, Joan Havel, and Bernard Seidenberg. "Is the F Scale Irreversible?" Journal of Abnormal and Social Psychology, 1958, 56, 143-159.

Cronbach, Lee J. "Further Evidence on Response Sets and Test Design." Educational and Psychological Measurement, 1950, 10, 3-31.

Cronbach, Lee J. "Response Sets and Test Validity." Educational and Psychological Measurement, 1946, 6, 475-494.

Cronbach, Lee J. "Studies of Acquiescence as a Factor in the True-False Test." Journal of Educational Psychology, 1942, 33, 401-415.

Garrett, Henry E. Statistics in Psychology and Education. New York: David McKay Company, 1958.

Leavitt, Harold J., Herbert Hax, and James H. Roche. "'Authoritarianism' and Agreement with Things Authoritative." The Journal of Psychology, 1955, 40, 215-221.

Mogar, Robert E. "Three Versions of the F Scale and Performance on the Semantic Differential." Journal of Abnormal and Social Psychology, 1960, 60, 262-265.

Ong, J. S. Louis. Measurement with Opposite Forms of an Inventory. Doctoral dissertation, University of Minnesota, 1963.

Peabody, Dean. "Attitude Content and Agreement Set in Scales of Authoritarianism, Dogmatism, Anti-Semitism, and Economic Conservatism." Journal of Abnormal and Social Psychology, 1961, 63, 1-11.

Rokeach, Milton. "The Double Agreement Phenomenon: Three Hypotheses." Psychological Review, 1965, 70, 304-309.

Rorer, Leonard G. The Function of Item Content in MMPI Responses. Doctoral dissertation, University of California at Berkeley, 1963.

Senesh, Lawrence. Our Working World: Families at Work. Chicago: Science Research Associates, Inc., 1964.

Shaver, James P. and A. Guy Larkins. Evaluation Report: SRA Economics Materials in Grades One and Two. (Mimeographed report, Utah State University, Logan, Utah, 1966.)